

Models of Protein Evolution

Lev Yampolsky* and Arlin Stoltzfus

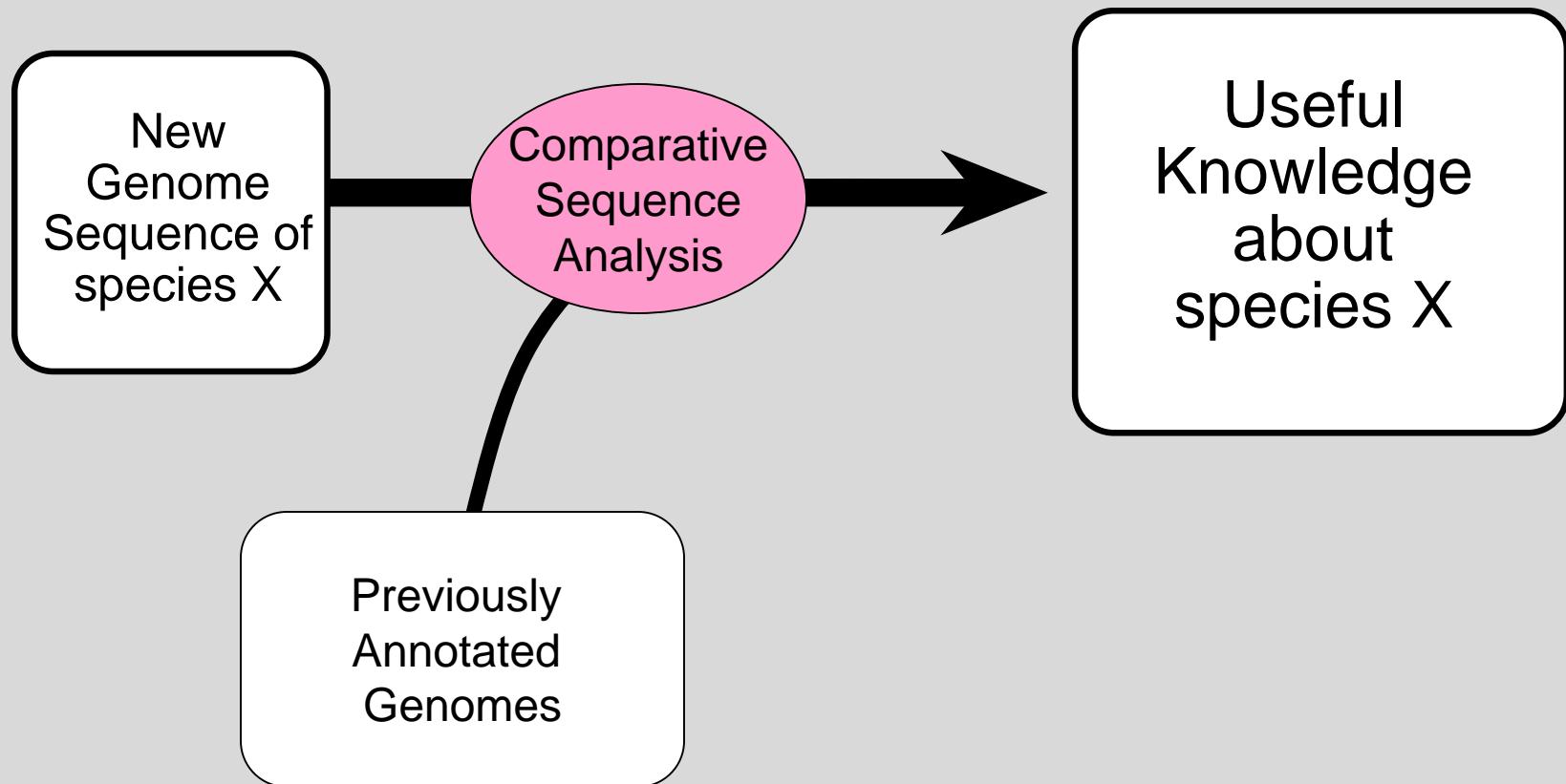
Center for Advanced Research in Biotechnology
Rockville, MD 20850

National Institute of Standards and Technology
Gaithersburg, MD

University of Maryland Biotechnology Institute
Baltimore, MD

*current: Department of Biology, East Tennessee State University,
Johnson City, TN

Computer-based analysis of genome sequence data



Advancing evolutionary bioinformatics

- Make evolutionary approaches more automatable
 - System for Phyloinformatic Analysis (SPAN)
(Qiu, Schisler & Stoltzfus)
- Improve theoretical models of evolution
 - Mutation-biased evolution of genes and proteins
(Yampolsky & Stoltzfus, 2001; Nawrocki & Stoltzfus, in prep.; Stoltzfus, 1999)
- Introduce parameter values for predictive models
 - Amino Acid Exchangeability from Experimental Data
(Yampolsky & Stoltzfus, 2003)
 - Mutation Parameters Estimated from Nonsense Mutants in Humans (Qiu & Stoltzfus, unpubl.)

Protein evolution as residue replacement

Assumptions of most protein evolution models

- Protein evolution is amino acid sequence evolution
- Sequence evolution occurs by replacing one residue with another

Typical assumptions of residue replacement models

- replacements at each site occur independently
- replacement is a first-order Markov process
- replacement is a time-reversible process
- replacement rates are identically distributed among sites

	Ala	Cys	Asp	Glu	Phe	...
Ala		α	β	χ	δ	
Cys	ε		ϕ	γ	η	
Asp	ι	φ		κ	λ	
Glu	μ	ν	σ		π	
Phe	θ	ρ	σ	τ		
...						

Asymmetric

	Ala	Cys	Asp	Glu	Phe	...
Ala		α	β	χ	δ	
Cys			ϕ	γ	η	
Asp				κ	λ	
Glu					π	
Phe						
...						

Symmetric

	T	C	A	G
T		$\mu\alpha$	μ	μ
C	$\mu\alpha$		μ	μ
A	μ	μ		$\mu\alpha$
G	μ	μ	$\mu\alpha$	

2-parameter (transition and transversion) model of nucleotide substitutions

The canonical genetic code

		Second Position Nucleotide				
		U	C	A	G	
U	Phe	UUU UUC	Ser	UCU UCC	Tyr UAU UAC	Cys UGU UGC
	Leu	UUA UUG		UCA UOG	STOP UAA UAG	STOP UGA Trp UGG
	C	CUU CUC CUA CUG	Pro	CCU CCC CCA COG	His CAU CAC	CGU CGC
	Leu				Gln CAA CAG	Arg CGA CGG
	A	AUU AUC AUA	Thr	ACU ACC ACA AOG	Asn AAU AAC	Ser AGU AGC
	Met	AUG			Lys AAA AAG	Arg AGA AGG
	G	GUU GUC GUA GUG	Ala	GCU GCC GCA GCG	Asp GAU GAC	GGU GGC
	Val				Glu GAA GAG	Gly GGA GGG

A mechanistic codon replacement model

$$R_{AAC \rightarrow AGC} = k f_{AAC} \mu_{A \rightarrow G} \pi_{Asn \rightarrow Ser}$$

But what is the basis for the acceptance function, π ?

- PAM (evolutionary transition probabilities)?
- BLOSUM (sequence conservation)?
- physicochemical distances (difference in volume, hydrophobicity, etc)?
- structural models of exchange effects (Miyazawa-Jernigan matrix)?

- Results of protein engineering experiments?
 - are data sufficiently systematic?
 - can results from diverse studies be combined?
 - can a measure of exchangeability be validated independently?
 - will the laboratory results prove relevant to phenomena in nature?

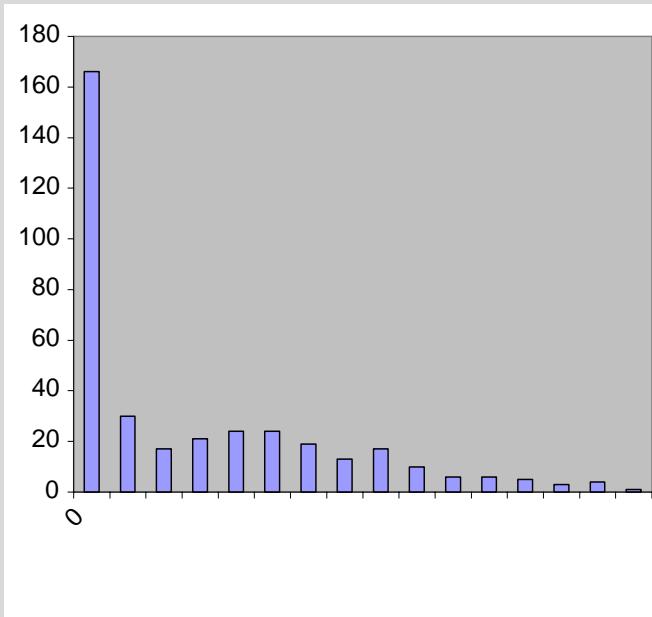
Summary of Studies

Protein	Method	Sites		Assay of effect	Reference
		(total)	Exchanges		
Lac Repressor	sup	328 (360)	4048	Lac operon repression, in vivo	Suckow et al., 1996
Lysozyme	sup	128 (164)	1581	plaque size, in vivo	Rennell et al., 1991
Interleukin-3	sat	103 (152)	754	cell proliferation activity, in vivo	Olins et al., 1995
Barnase	sat	109 (110)	676	RNAse activity, in vivo	Axe et al., 1998
b-Lactamase	sat	27 (246)	513	ampicillin resistance, in vivo	Palzkill et al., in prep.
RecA	sat	20 (323)	380	lambda plaque assay, in vivo	Hortnagel et al., 1999
RTase	sat	109 (300)	366	RNA-dep. DNA pol. activity, in vitro	Wrobel et al., 1998
Protease	sat	99 (99)	336	protease activity, in vivo	Loeb et al., 1989
Gene V protein	sat	86 (87)	313	inhibition of E.coli growth, in vivo	Zabin et al., 1991
Nuclease	scan	143 (149)	290	thermodynamic stability, in vitro	Shortle et al., 1990, 1992, 1996
HGH	scan	50 (191)	50	Kd for receptor, in vitro	Cunningham and Wells, 1989
Insulin	scan	37 (51)	37	affinity to receptor, in vitro	Kristensen et al., 1997
Total		1239	9334		

Numbers of Exchanges

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	Sum
C	7	2	7	7	9	1	1	7	6	6	6	8	6	3	1	7	2	8	9	2	99
S	44		13	44	60	45	8	5	39	38	36	43	39	4	7	41	5	41	43	3	558
T	30	53		51	57	41	16	7	31	32	32	40	36	5	21	31	9	29	30	5	556
P	20	32	21		44	29	5	6	21	25	28	38	25	5	6	40	5	21	19	6	396
A	51	67	16	70		79	5	12	58	52	53	55	54	4	2	60	25	52	53	3	771
G	57	59	15	46	77		11	24	53	45	43	63	48	10	14	45	43	42	43	13	751
N	23	35	17	27	36	30		16	23	25	36	27	32	7	17	31	9	24	31	7	453
D	28	31	7	32	56	55	24		51	31	44	31	31	5	5	33	25	30	47	4	570
E	21	30	6	27	54	45	5	15		28	23	25	29	4	3	28	13	21	24	3	404
Q	29	35	5	44	42	35	3	1	42		40	41	35	4	5	44	5	30	28	5	473
H	12	14	7	17	21	18	11	8	13	18		19	14	7	7	19	6	16	17	6	250
R	35	45	9	39	43	44	3	4	30	32	40		36	1	8	42	6	30	29	3	479
K	21	34	24	25	51	46	22	5	34	41	24	42		15	14	27	7	22	23	7	484
M	14	13	3	14	21	18	4	3	13	13	14	16	15		7	17	5	14	13	2	219
I	30	50	27	35	42	38	17	4	33	31	31	44	35	23		58	30	47	28	4	607
L	55	74	14	72	86	73	10	7	61	68	63	79	57	15	19		33	68	56	14	924
V	46	48	7	47	69	67	4	11	51	45	47	47	44	5	17	57		53	44	4	713
F	21	21	4	15	23	15	4	4	13	11	13	14	14	2	13	24	11		21	4	247
Y	30	27	0	16	26	22	11	12	14	14	26	18	14	1	0	16	0	31		2	280
W	11	8	1	6	6	12	0	0	5	6	5	9	6	1	0	11	1	6	6	6	100
Sum	578	683	198	634	821	721	164	145	592	561	604	659	570	121	166	631	240	585	564	97	9334

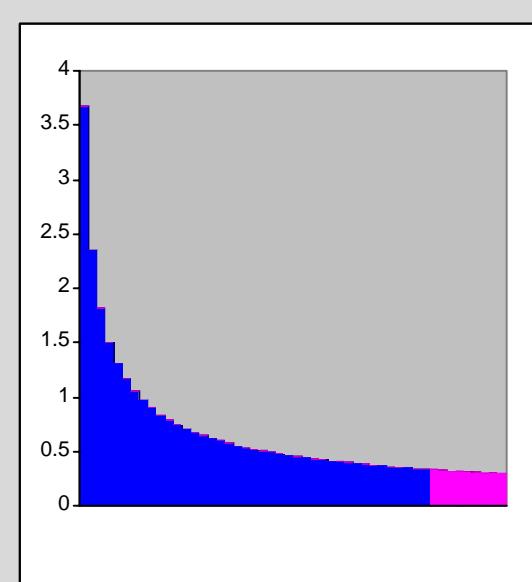
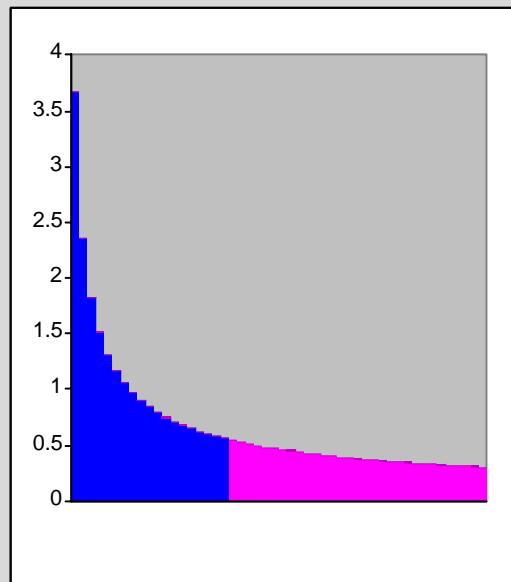
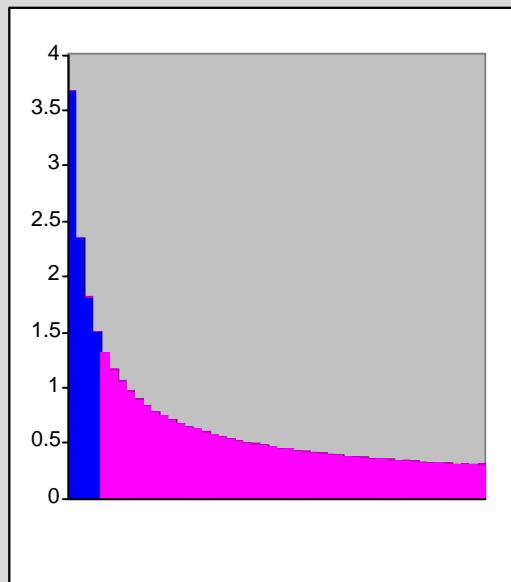
Problem: how to combine results?



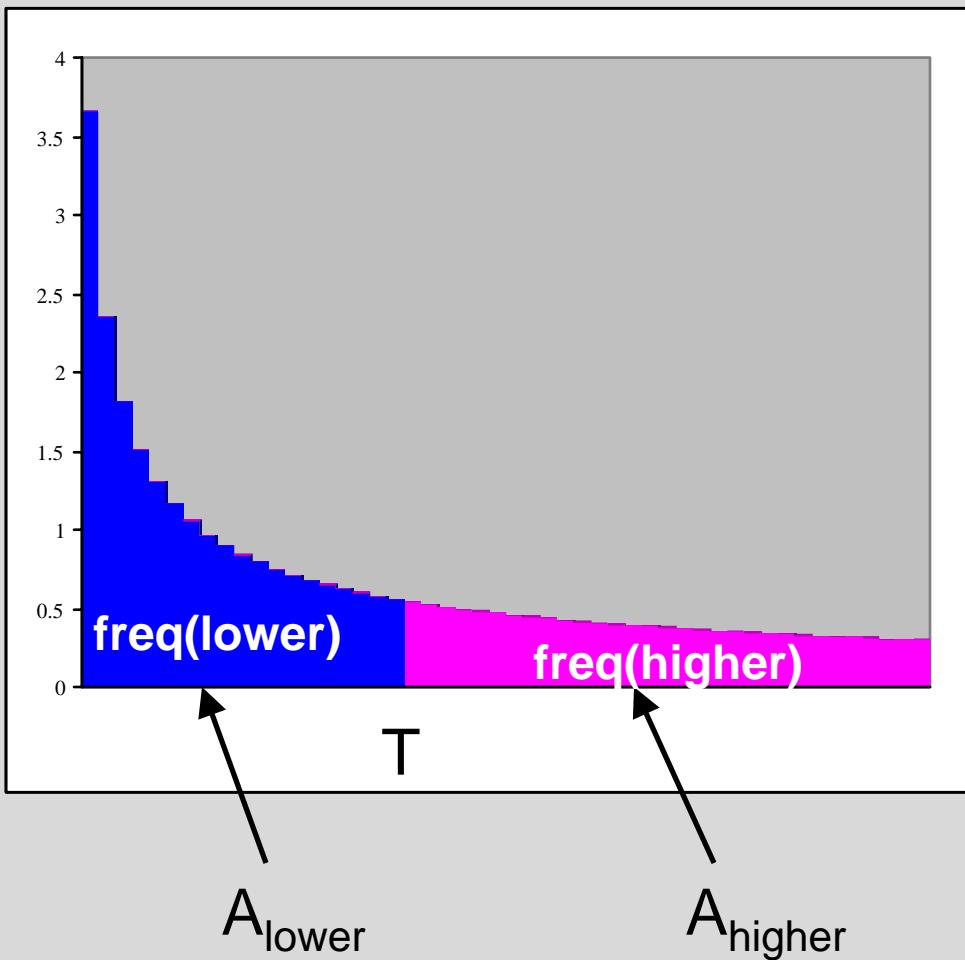
Protein	% Loss	% Retain	Basis of Loss/Retention Determination
Barnase	4.9	95	Cell-killing assay of RNase activity requires about 0.2% wild-type.
Lysozyme	20	80	Ability of T4 to form normal plaques requires about 3% wild-type.
b-Lactamase	93	7.2	Ability to confer ampicillin resistance requires about 50% wild-type.

Problem: how to combine results?

Protein	% Loss	% Retain	Basis of Loss/Retention Determination
Barnase	4.9	95	Cell-killing assay of RNase activity requires about 0.2% wild-type
Lysozyme	20	80	Ability of T4 to form normal plaques requires about 3% wild-type
β -Lactamase	93	7.2	Ability to confer ampicillin resistance requires about 50% wild-type

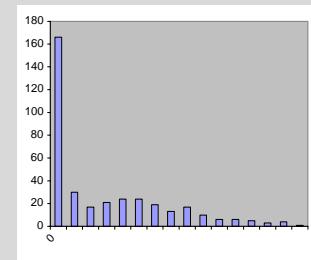
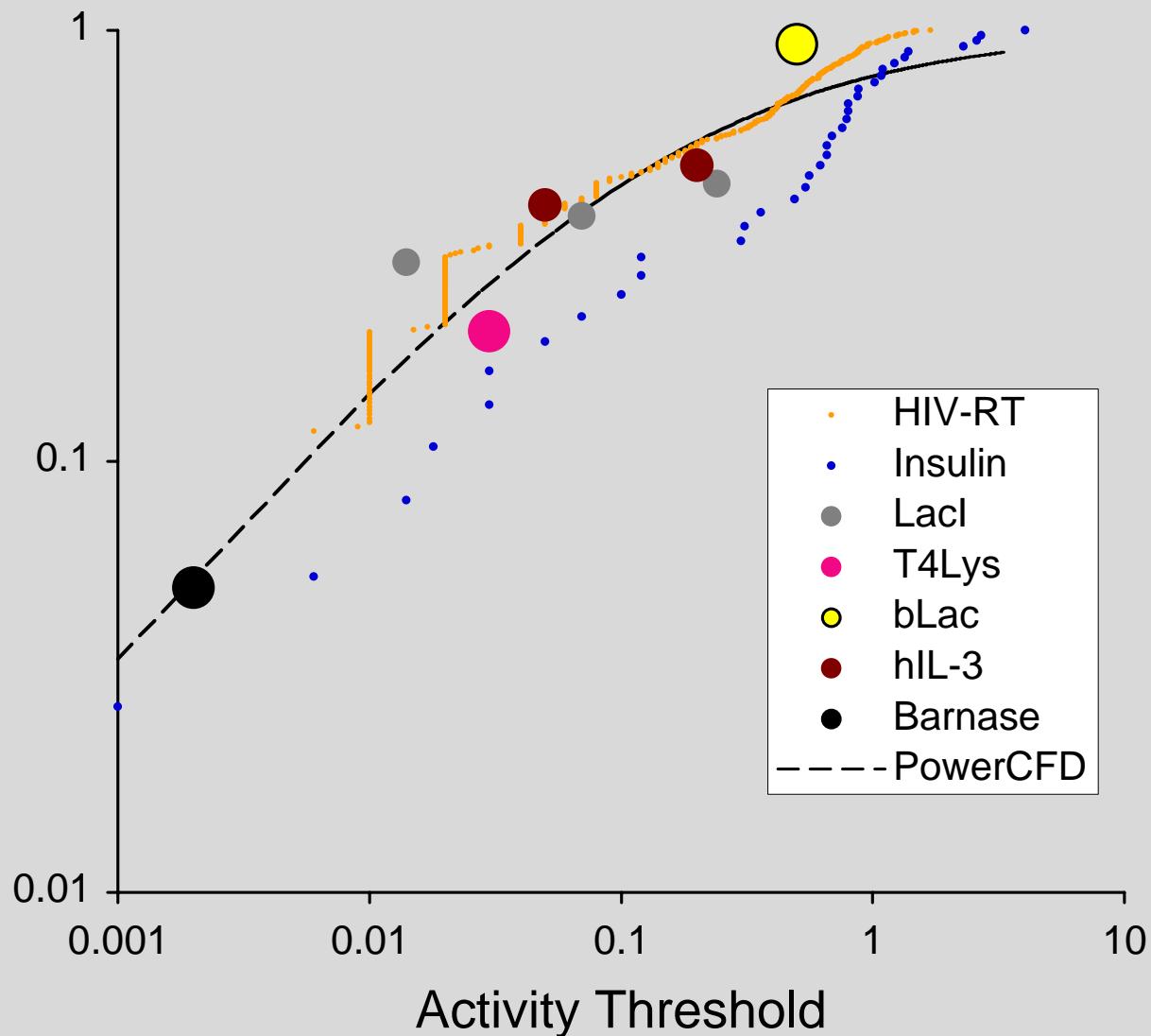


Transforming scores



- Compute threshold T from known category frequencies
- Assign scores for categories from known thresholds

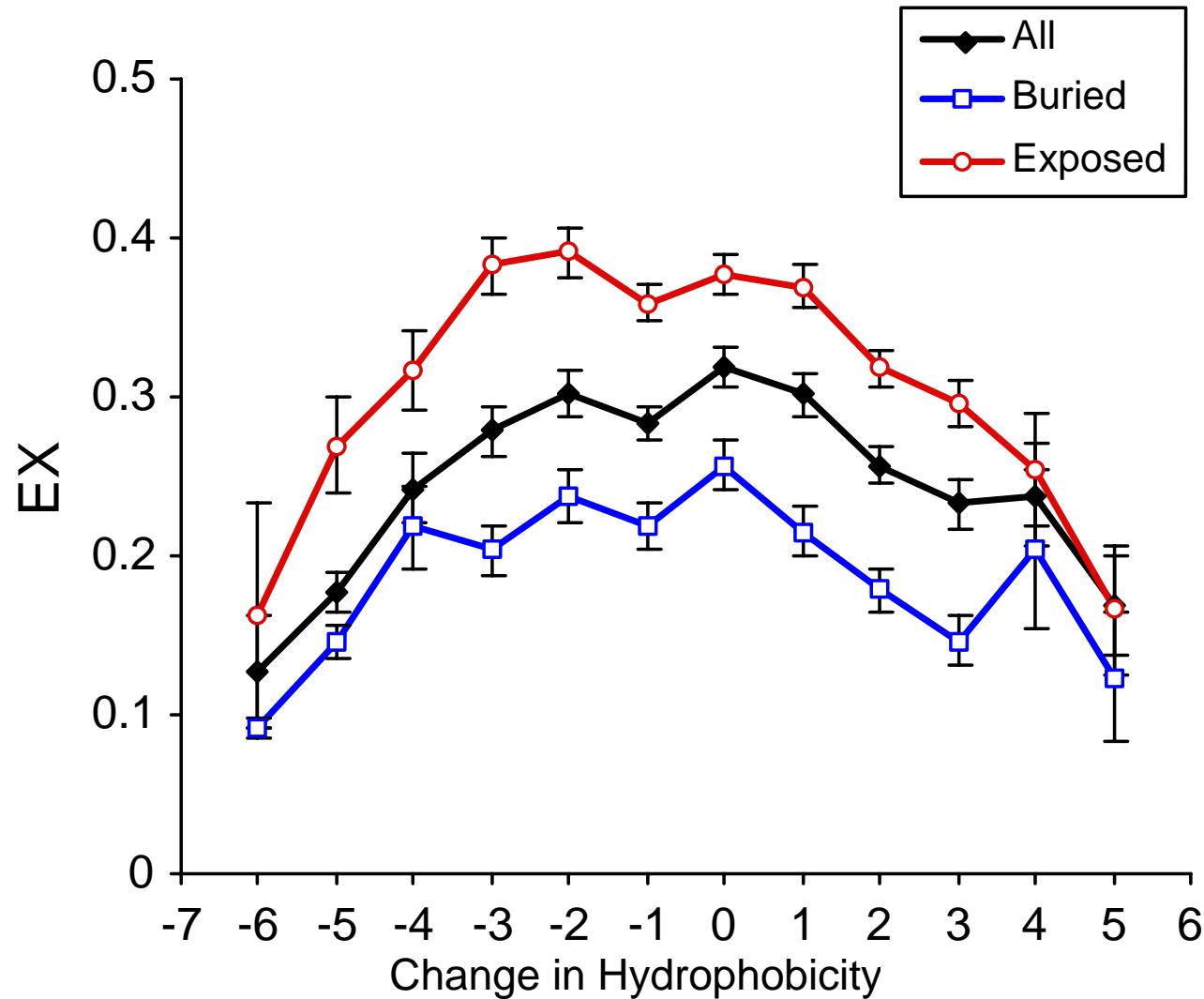
Frequency impaired vs. threshold



Asymmetric Exchangeability

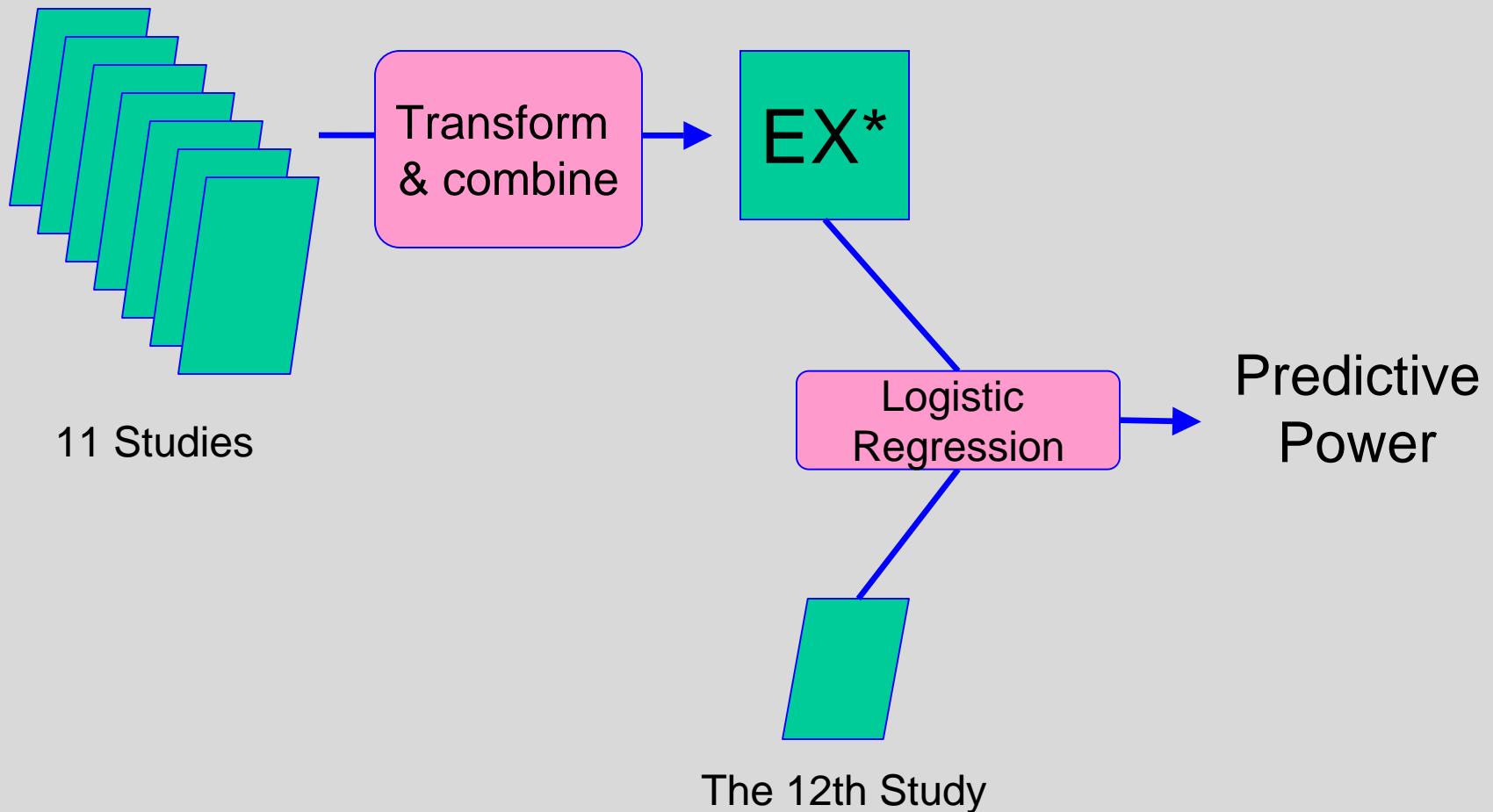
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	0.39	0.13	0.28	0.46	0.36	0.14	0.14	0.39	0.52	0.35	0.40	0.34	0.09	0.14	0.46	0.08	0.47	0.45	0.23	
S	0.45		0.46	0.26	0.52	0.48	0.44	0.38	0.38	0.41	0.40	0.42	0.34	0.34	0.30	0.32	0.29	0.38	0.35	0.25
T	0.38	0.50		0.19	0.44	0.37	0.31	0.26	0.23	0.32	0.27	0.34	0.25	0.21	0.28	0.32	0.37	0.31	0.29	0.09
P	0.45	0.43	0.32		0.45	0.42	0.39	0.33	0.45	0.48	0.42	0.30	0.27	0.33	0.21	0.31	0.45	0.37	0.35	0.35
A	0.49	0.49	0.38	0.29		0.45	0.38	0.35	0.31	0.37	0.36	0.33	0.26	0.58	0.28	0.39	0.25	0.37	0.34	0.23
G	0.34	0.40	0.26	0.18	0.41		0.32	0.26	0.26	0.34	0.30	0.27	0.26	0.31	0.16	0.26	0.21	0.22	0.24	0.25
N	0.33	0.44	0.30	0.31	0.42	0.45		0.27	0.32	0.38	0.31	0.32	0.23	0.24	0.26	0.28	0.25	0.27	0.32	0.15
D	0.34	0.34	0.23	0.25	0.33	0.31	0.25		0.40	0.35	0.35	0.30	0.23	0.19	0.30	0.27	0.23	0.26	0.29	0.11
E	0.52	0.48	0.32	0.25	0.52	0.45	0.18	0.50		0.49	0.52	0.46	0.48	0.17	0.27	0.47	0.33	0.52	0.52	0.14
Q	0.47	0.52	0.29	0.28	0.52	0.48	0.39	0.10	0.50		0.50	0.48	0.44	0.47	0.44	0.48	0.52	0.45	0.47	0.18
H	0.33	0.44	0.24	0.23	0.47	0.40	0.26	0.18	0.32	0.39		0.31	0.33	0.34	0.24	0.44	0.27	0.32	0.29	0.10
R	0.27	0.32	0.22	0.19	0.44	0.29	0.08	0.06	0.30	0.38	0.32		0.39	0.10	0.13	0.32	0.22	0.28	0.34	0.07
K	0.43	0.42	0.43	0.29	0.51	0.47	0.43	0.33	0.34	0.48	0.46	0.54		0.41	0.46	0.39	0.41	0.46	0.45	0.15
M	0.46	0.43	0.48	0.09	0.40	0.25	0.51	0.45	0.33	0.48	0.34	0.15	0.15		0.54	0.62	0.55	0.47	0.36	0.71
I	0.43	0.26	0.28	0.17	0.37	0.19	0.24	0.07	0.23	0.24	0.28	0.14	0.13	0.34		0.50	0.50	0.40	0.41	0.09
L	0.48	0.27	0.18	0.19	0.39	0.25	0.21	0.13	0.25	0.33	0.35	0.23	0.21	0.42	0.32		0.33	0.43	0.38	0.14
V	0.41	0.36	0.47	0.22	0.41	0.28	0.13	0.38	0.18	0.32	0.28	0.21	0.21	0.48	0.55	0.35		0.26	0.25	0.28
F	0.24	0.23	0.42	0.14	0.29	0.13	0.22	0.08	0.09	0.24	0.30	0.14	0.11	0.42	0.31	0.37	0.39		0.41	0.39
Y	0.23	0.24	ND	0.17	0.35	0.31	0.20	0.14	0.24	0.34	0.28	0.29	0.24	0.13	ND	0.38	ND	0.41		0.33
W	0.19	0.15	0.01	0.10	0.10	0.25	ND	ND	0.11	0.09	0.33	0.17	0.09	0.11	ND	0.30	0.11	0.47	0.38	

Exchangeability vs. Δ Hydrophobicity



Statistical Cross-validation

(a method to validate EX)



Comparative evaluation

Study	Size	Predictor								
		EX	EXS	VB	BLOSUM	PAM250	XX	lnMJ	1/Ga	
Lac Repressor	4038	9.1E-52	2.8E-14	7.7E-27	6.5E-33	2.4E-29	4.5E-35	1.1E-16	9.4E-07	
T4 Lysozyme	1918	1.1E-41	1.8E-16	5.2E-16	1.3E-17	1.6E-15	7.0E-15	3.3E-12	2.3E-04	
Interleukin-3	754	5.7E-11	7.2E-11	1.7E-10	2.2E-08	2.5E-10	7.8E-10	1.4E-08	2.5E-03	
Barnase	676	5.9E-07	1.0E-06	5.7E-04	1.8E-04	1.0E-03	3.0E-04	7.8E-02	5.0E-02	
b-Lactamase	513	4.3E-02	7.5E-03	7.0E-02	8.1E-03	7.5E-03	9.5E-02	2.7E-04	4.6E-01	
RecA	380	2.8E-03	7.7E-03	2.1E-03	1.1E-03	2.5E-03	7.1E-03	7.1E-01	1.4E-04	
HIV RTase	366	2.5E-22	2.5E-15	3.6E-09	8.4E-16	3.3E-14	4.6E-11	1.2E-05	2.8E-05	
HIV Protease	336	1.9E-16	1.2E-05	3.0E-09	1.1E-17	3.5E-15	2.3E-16	1.4E-07	4.1E-06	
f1 Protein V	313	5.1E-08	3.6E-06	1.4E-03	1.8E-07	3.3E-06	7.1E-06	2.4E-03	7.9E-02	
Staph. Nuclease	290	8.7E-11	9.4E-04	4.7E-03	7.0E-06	1.4E-11	1.3E-03	5.8E-04	1.1E-02	
HGH	50	3.6E-01	4.5E-01	9.8E-02	4.0E-01	2.2E-01	3.1E-01	1.6E-01	6.7E-01	
Insulin	37	2.0E-02	2.8E-02	5.5E-02	1.1E-01	1.6E-02	4.4E-01	1.6E-01	8.9E-01	
All	9671	1.6E-75	7.6E-47	2.1E-46	5.8E-44	3.4E-37	4.4E-32	3.9E-21	1.8E-06	

EX as the basis for an acceptance function

$$(\text{reminder: } R_{AAC \rightarrow AGC} = k f_{AAC} \mu_{A \rightarrow G} \pi_{Asn \rightarrow Ser})$$

Log likelihood from a codon-based model
of the evolution of 12 mitochondrial protein-
coding genes from 7 primate species.

Yang, Nielsen and Hasegawa, 1998.

Results from this study

Distance Measure	Transform	p $\tilde{\Theta}$	lnL
None (equal D)	NA	11	-29967.86
Composition	geometric	12	-29961.86
Composition	linear	12	-29961.34
Polarity	geometric	12	-29917.48
Polarity	linear	12	-29917.34
Volume	geometric	12	-29900.76
Volume	linear	12	-29909.83
Grantham's D	geometric	12	-29911.54
Grantham's D	linear	12	-29915.68
Miyata's D	geometric	12	-29890.18
Miyata's D	linear	12	-29895.91
1 - EXS	geometric	12	-29885.81
1 - EXS	linear	12	-29871.73

Summary

- Experimental exchange studies are a rich source of systematic data
- Results of diverse studies can be combined
- Comparative evaluation shows that EX out-performs other distance measures with respect to laboratory effects of exchanges
- With respect to natural amino acid exchanges, EX also seems to improve upon available measures

Acknowledgements

Phyloinformatics of Introns

- WeiGang Qiu
- Arlin Stoltzfus
- Nick Schisler (Pomona College)
- Eric Nawrocki

Amino acid exchangeability

- Lev Yampolsky
- Arlin Stoltzfus

Duplicate gene evolution

- Oisin Feeley
- Arlin Stoltzfus

Mutation biases (theory/simulation)

- Arlin Stoltzfus
- Lev Yampolsky
- Eric Nawrocki

Mutation biases (HGMD analysis)

- Weigang Qiu
- Lev Yampolsky